DOCUMENT RESUME

ED 249 270 TM 840 623

AUTHOR Alliger, R. J.; Harvey, A. L.
TITLE Problems in Measuring Formal Operations.
PUB DATE Apr 84
NOTE 29p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 24-26, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Developmental Stages; Group Testing; *Interviews; *Objective Tests; Scoring; *Testing Problems; Test Reliability; Test Validity
IDENTIFIERS *Formal Operations; Paper and Pencil Tests; *Piagetian Stages

ABSTRACT

This article discusses practical and theoretical problems related to the measurement of formal operations. The first section of the article discusses problems in measuring formal operations using the clinical interview method. These problems include the lack of both a standardized interview and a uniform scoring procedure. Section two discusses some of these problems which may be solved by utilizing group-administered paper-and-pencil tests of formal operations. Section three discusses problems with paper-and-pencil formal operations tests. The major focus in this section is on problems related to the measurement of reliability and validity. Problems with response constraint and reading comprehension level are also discussed briefly. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED249270

Problems in Measuring Formal Operations

R. J. Alliger A. L. Harvey

University of Nebraska-Lincoln

Paper presented at the Annual Meeting of the National Council on Measurement in Education. New Orleans, 1984

Running Head: PROBLEMS IN MEASURING FORMAL OPERATIONS

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

R J. Alliger

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 840623

Abstract

This article discusses practical and theoretical problems related to the measurement of formal operations. The first section of the article discusses problems in measuring formal operations using the clinical interview method. These problems include the lack of both a standardized interview and a uniform scoring procedure. Section two discusses some of these problems which may be solved by utilizing group-administered paper-and-pencil tests of formal operations. Section three discusses problems with paper-and-pencil formal operations tests. The major focus in this section is on problems related to the measurement of reliability and validity. Problems with response constraint and reading comprehension level are also discussed briefly.

Problems in Measuring Formal Operations

The popularity of formal operations as a variable in educational research and clinical evaluations is increasing steadily. Along with this rapid growth has come misuse which requires serious attention from both developmentalists and psychometricians. Both theoretical and measurement problems exist for the researcher, clinician, or instructor wishing to determine the developmental level of a group of individuals. While many of these problems apply to all of Piaget's developmental levels, the discussion here will be restricted to the problems involved in the measurement of formal operations. Many of the tests of formal operations developed recently have been inadequately and incorrectly evaluated before their use in research and clinical settings. The intention of this paper is to point out these concerns so a more controlled measurement progress for these tests may be accomplished.

Piaget's clinical interview method has long been popular among developmentalists. The interview allows the researcher flexibility in both format and direction. As pointed out in Brainerd (1978), it does not appear that Piaget intended his interviews to be rigid and standardized measurement techniques. Piaget's method was helpful in describing a pattern of thought development within the individual, not in measuring thought relative to a group of individuals. While Piaget was interested

in looking at how thought progressed, researchers are often attempting to pinpoint a particular place in that progression. Many researchers wish to compare an individual's performance (relative to the group) with performance on a continuous variable. Piaget's interviews may be useful in pointing out how thought changes while solving a specific type of task (e.g. the balance task), but the method is not intended to unequivically categorize an individual as "concrete operational" or "formal operational" for all tasks. It is, therefore, not particularly useful in ranking one individual as more or less formal operational than another.

The following sections will explore how this difference in approach affects the usefulness of the interview technique in research. This brief discussion will be followed by two sections which explore the relative advantages and disadvantages of the increasingly popular paper-and-pencil exams used to measure level of cognitive development from a Piagetian perspective.

These sections are designed primarily as an overview for use by educators and researchers outside the areas of measurement and/or development. For a more in-depth discussion, see Nagy & Griffiths (1982), Lunzer (1965), and Gray (1978), among others.

The Piagetian Interview

Probably the most obvious and most frustrating of the problems in measuring formal operations is the problem of horizontal

decalage. Research has shown the same individual to vary in achievement of formal operations across type of problem, for example, proportional reasoning vs. probabilistic reasoning (Martarano, 1977; Tschopp & Kurdek, 1981), and also across problem domains, for example, from English to mathematics (Higgins-Trenk & Gaite, 1971; Piaget, 1972; Wollman & Karplus, 1974). Coupled with the evidence that individuals may revert to earlier levels of functioning in times of stress (Breger, 1974), leads to the conclusion that formal operations may be a very elusive and unstable construct. Reliability is then, at best, difficult to establish since the individual's responses are unreliable.

The instability and domain specificity of formal operations necessitates great care in the reporting and interpreting of research results. Most researchers use problems from the domains of the sciences in establishing formal operations. If formal operations is, indeed, domain specific this selective measurement limits the generalizability of the results.

Results may also be limited in another way. The interview format assumes that individuals being tested are capable of expressing themselves verbally and are willing to do so. This is obviously not always the case. A skilled interviewer must therefore conduct the session. The interviewer must be capable of a) establishing good rapport, b) motivating the subject to communicate what may be a difficult concept (i.e. a concept

beyond his or her present vocabulary), and c) discerning the difference between a lack of understanding and a lack of vocabulary.

The interviewer must also motivate the subject to find the formal operational solution to the task. When the correct answer to a task is possible using concrete operations many subjects may opt for this solution rather than make the extra effort to find the "best" solution (Higgins-Trenk & Gaite, 1971).

The Piagetian interview, originally intended as a research technique, is perhaps one of the most unstandardized measurement tools on record. Different researchers may not just use different tasks but may present the tasks in different orders using different equipment. Further, one of the trademarks of Piagetian techniques is its flexibility during the interview. While particular researchers may have standardized the technique for their own use, this standardized version is very often not used by others investigating the same topic. A limit to the generalizability of results occurs due to this lack of consistency. While it may be that a group of researchers finds their own interview technique to be quite reliable, a similar procedure conducted by someone outside their sphere of influence may show far different results. Without a standard interview with standard problem tasks using standard equipment, unreliability is quite likely.

Closely related to the problem of an unstandardized format is the lack of a uniform scoring system. Piaget's own scoring system, while elusive to anyone but an expert, may be used for individual problems. No uniform scoring system exists, however, for the scoring or interpretation of a series of tasks. How many tasks must be completed at the formal operational level for an individual to be considered formal operational? Or will an average be used? Where are the criteria for cutoffs?

While the previous problems are pervasive, the most noticeable problem to the individual researcher is the technique's time consuming nature. Researchers and instructors who want to test large numbers of subjects often find the prospect of long individual interviews daunting. In response to this problem, and to others mentioned, a veritable potpourri of group tests measuring formal operations have been developed. The following sections will explore the advantages and disadvantages of using these tests.

Advantages of Group Tests

Paper-and-pencil tests of formal operations have flourished in the last several years. Even a casual browse through related literature will turn up over a dozen, and its probably safe to assume this is but the "tip of the iceberg". A characteristic of this proliferation is the diversity in methods. Tests appear to differ along some key dimensions that could potentially affect

their usefulness: a) whether or not demonstrations are used, and if used, via the instructor or by videotape, b) whether the examinees are asked to supply written reasons for their choice of answer, or are asked to choose from supplied reasons in a multiple choice format, and c) whether the test is scored according to individual questions or statements exhibiting varying degrees of concrete or formal thought, or whether it is scored according to correct or incorrect answers.

By way of example, both the cognitive examination developed by Tomlinson-Keasey & Campbell (unpublished) and the Test of Logical Thinking (TOLT) developed by Tobin and Caple (1981) present a relationship problem involving mice. Both ask the examinee to state whether or not there is a relationship. The TOLT then asks the examinee to pick from among five reasons why they chose their answer. The Tomlinson-Keasey & Campbell test, however, requires a written explanation of the examinees choice. The item would then also be scored differently for each of the examinations. The Tomlinson-Keasey & Campbell item would be scored according to the level of concrete or formal thought expressed, similar to Piaget's system. The TOLT item would simply be scored correct or incorrect.

This diversity of testing procedure leads to a variety of advantages over the interview technique, depending on the type of written test. The advantages they all have in common are that of allowing group administration and allowing that administration to

proceed in a standardized manner. The need for group tests is readily apparent, particularly from a research perspective. From a psychometric perspective, it would appear that the standardization aspect is an equally appealing improvement.

The advantages of a standardized instrument are obvious: 1) the quality of the instrument is now testable, 2) results can be compared from one situation to another, 3) informed decisions can be made as to the accuracy of results, and 4) the level of expertise needed to use the instruments outside a clinical setting is lowered considerably from that of the interview method.

In addition, some of the instruments may offer other advantages. For instance, those involving multiple choice items may eliminate the problem of examinees being unable to express their solutions to a task. Most eliminate, to a large extent, the obtrusiveness of the interviewer. The results are not dependent on the expertise of the interviewer in motivating and directing the examinee without influencing the natural outcome of the subject's thought processes. Finally, many of the instruments allow easy scorability by non-experts, as well as more reliable scoring by all.

These advantages have lead to the popularity and increased use of standardized instruments, but not without mixed results. The comparative ease in using an instrument of this type has lead to the indiscriminate development and dissemination of relatively

unsupported instruments. Many of these instruments are widely used after only the barest minimum of quality testing. This is further compounded by the fact that developmental testing is hampered by its own, unique problems which disallow the use of most traditional test analysis techniques. The following section explores those problems unique to paper-and-pencil tests of formal operations.

Problems With Paper-and-Pencil Tests

The two most universal concepts in measurement are reliability and validity. Measurement problems with paper-and-pencil formal operations tests surrounding reliability will be discussed first followed by a discussion of validity issues.

Reliability

Test-retest reliability. Test-retest reliability is not usually appropriate for formal operations tests. Subjects retested after a short period of time may be influenced by practice effects. Nagy and Griffiths (1982) note that initial testing with group tests (i.e., paper-and-pencil test) may act as a teaching device and raise scores on the second testing. Such a practice effect will likely be different across students (Mehrens & Lehmann, 1975) and, consequently, lower reliability. Memory effects may also influence test-retest reliability over a short period of time. Subjects may mark an answer on the second testing not because they think it is the correct answer, but

because it is the answer they remember marking on the first test. This would make test-retest reliability spuriously high. Problems in using test-retest reliability with formal operations tests also exist over long retest periods of time. Scores of subjects retested after a long period of time may change due to developmental growth (Nagy & Griffiths, 1982). Lower reliability indices may therefore result.

Parallel forms reliability. Parallel forms reliability is also an inappropriate measure of reliability for formal operations tests. Formal operations is not an all-or-none phenomenon (Tschopp & Kurdek, 1981). A subject who uses formal reasoning on one task may not use it on another task. These intraindividual differences (Martorano, 1977) render parallel forms reliability useless because it is virtually impossible to determine equivalent test items/tasks. What appears as a parallel item to an investigator may be a totally different picture to a subject (Nagy & Griffiths, 1982).

Internal consistency reliability. Reliabilities reported for formal operations paper-and-pencil tests are generally internal consistency measures and are generally quite acceptable (see Lawson, 1978; Staver & Gabel, 1979; and Tobin & Capie, 1981). There are however, problems in using internal consistency measures of reliability.

Two aspects of the measurement of formal operations, scaling problems and the criterion-referenced nature of development tests, call into question the use of internal consistency measure of reliability.

The goal of Piagetian research methodology is often to determine the qualitative cognitive differences which explain an individual's behavior (Modgil & Modgil, 1976). Formal operations tests have focused on the quantitative measurement of individual behavior. This difference reflects a difference in measurement scales utilized by the two methodologies. (NOTE: The two methodologies also utilize different dependent variables, but this will be discussed later in the validity section). Piagetian methodology utilizes an ordinal level measurement scale while paper-and-pencil tests of formal operations often utilize an interval level measurement scale. Interval level data is required in order to calculate internal consistency measure of reliability. In this sense, formal operations tests meet the necessary measurement requirements. The problem is the measurement scale used by formal operations tests is different than intended by the theory. Separating the measurement scale from the measurement theory can result in different interpretations of outcomes.

Uzgiris and Hunt (1975) note conceptual differences between ordinal measurement scales and traditional measurement scales related to the organization and development of intelligence.

Traditional tests of intelligence measure intelligence in an incremental manner. Little consideration is given to the relationship between achievement at one level and that at the next level. Ordinal scales imply a hierarchical relationship between achievement at different levels. Achievements reached at one level are not possible without having first achieved the previous level. Elkind (1981) provides a very useful illustration of the difference between Piagetian and traditional tests of intelligence.

> In the Piagetian view, therefore, mental growth is not a quantitative but rather a qualitative affair and presupposes significant differences between the thinking of children and adolescents as well as between pre-school and school age children.
>
> These qualitative differences are, in fact, built into the items of mental tests but are obscured by the assignment of point scores to successes and failures. On the Wechsler Intelligence Scale for Children various of the sub-tests recognize qualitatively different responses only by assigning them additional points. A child, to illustrate, who says that a peach and a plum are alike because "they both have pits" is given a single point whereas a child who says "they are both fruit" is given two points. On other subtests, such as the arithmetic sub-test, there is

> no point differential success on problems which patently require different levels of mental ability . . . Clearly, the items on any given sub-test can tap quite different mental processes but these qualitative differences are ignored by assigning equivalent point scores to the various items regardless of the mental processes involved. (pp. 146-147)

An important difference between ordinal and interval scaling which Uzgiris and Hunt (1975) and Elkind (1981) do not directly point out is the lack of equal intervals in ordinal scales. Interval scales assume equal intervals between measurement points. Internal consistency measures of reliability assume the presence of equal intervals. Although it can be argued few tests actually are representative of equal intervals, this assumption is blatantly violated in Piagetian theory. Thus, to assume otherwise in formal operations tests is to conveniently ignore a theoretical given. This violation makes the use of internal consistency measures of reliability commonly reported highly questionable. As an example, one method of scoring formal operations tests involves scoring subjects right or wrong on test items and summing the total items correct (Arlin, 1982; Tobin & Capie, 1981). Essentially, this suggests that the higher a subject's score, the more they are formal operational. This is an incremental perspective of formal operational thought which assumes equal distances from increment

to increment. Piaget is in theory interested in qualitative differences in thought which were not necessarily equidistant.

Whether or not formal operations tests are utilizing an appropriate measurement scale is a problem which needs to be resolved. If it is appropriate to make a conceptual leap from ordinal to interval level measurement, internal consistency measures are perfectly appropriate. A greater threat to whether internal consistency reliability is an appropriate measurement index is that of the criteria used to compare subjects to each other.

Norm-referenced tests assess the performance of a subject by comparing it with those of other subjects' performance; i.e., the norm group (Mehrens & Lehmann, 1975). The focus is on interindividual variability. This perspective is in stark contrast to Piagetian theory. Hunt (1975) notes how the meaning of Piagetian tasks does not come from comparing a score to that of other subjects, but from the position of the score relative to a hypothesized universal developmental order.

> In Piagetian psychometry the tasks are therefore structured and interpreted according to a logic rather than a mere statistical comparison with other children. As such, tasks embedded in the Piagetian developmental theory have the long sought after character of immediately relevant criterion measures of development. This promising criterion-referenced

> character of Piagetian measures can be made clearer by brief theoretical and empirical comparisons of the traditional psychometric and Piagetian approaches to assessment. (Hathaway & Hathaway-Theunissen, 1974; cited in Hunt, 1975).

It seems apparent norm-referenced measurement is inappropriate for Piagetian measures. Criterion-referenced measurement, on the other hand, assesses a subject's score by comparing it to some specified criterion (Mehrens & Lehmann, 1975). A subject's score helps a researcher determine whether the subject has reached formal operational thought as opposed to the subject's level of cognitive development relative to other subjects. Gray (1978) has provided an in-depth discussion of the similarities as well as some of the differences between Piagetian theory and criterion-referenced measurement. It is noted how criterion-referenced measurement and Piagetian theory stress intraindividual variability (not interindividual variability as in norm-referenced testing). Formal operations tests would therefore seem to fit more appropriately into the category of a criterion-referenced test. How well subjects do on a formal operations test is judged relative to criterion levels specified by Piaget. They are not judged relative to all other subjects. Internal consistency measures of reliability do not apply to criterion-referenced tests and, therefore, are inappropriate for use with formal operations tests.

Validity

Content validity. Content validity involves defining a domain of behaviors to be covered by a test and then designing items which represent that domain (Allen & Yen, 1979). Piagetian theory and formal operations tests make it difficult to fulfill the requirements of the definition. First, formal operations is not defined in terms of behaviors. Piagetian theory is interested more in internal thought processes than observable behavioral manifestations of thought processes. Rowell & Hoffman (1975) state that replacing the interactive verbal interview with written responses and often multiple choice formats may very well change what is being measured. Second, as was noted earlier, it is difficult to establish priority of what type of tasks will elicit the presence of a general formal operational ability. Formal operational ability appears to be task specific (Nagy & Griffiths, 1982; Tschopp & Kurdek, 1981). This suggests research would be more fruitful if it were to develop formal operational tests specific to particular fields of study (Nagy & Griffiths, 1982).

Criterion-related validity. Criterion-related validity involves correlating the test scores of a paper-and-pencil formal operations test and an independent external measure of formal operations; i.e., subject scores from Piagetian interviews. Although some researchers have found acceptable criterion-referenced validity between formal operations test scores and Piagetian interview scores

(Lawson, 1980; Tobin & Capie, 1981), others report low correlation coefficients between the two measures (Bart, 1972; Blake, 1980; Staver & Gubel, 1979; Tschopp & Kurdek, 1981). These contradictory results may be in part due to the unstandardized nature of the Piagetian interview method. Some researchers may be more skilled interviewers and able to extract information needed to judge subjects more effectively. The high inter-rater reliabilities reported by investigators utilizing the Piagetian interview methodology (Rowell & Hoffman, 1975; Tschopp & Kurdek, 1981) would seem to refute this claim. However, high inter-rater reliabilities can be misleading. Piagetian interview methodology has a strong subjective element and if the investigator has weak interview skills, what they think is the subject's developmental level and what it actually is may be two separate realities. Further, it is intuitively apparent that investigators in a single study will have high inter-rater reliabilities because they, in all likelihood, work and train together. The inter-rater reliabilities may be quite different between investigators trained in two separate locations. If so, low inter-rater reliabilities could lower criterion-related validity because of disagreement in subject developmental level.

Construct validity. Construct validity is the degree to which a test measures a theoretical construct or trait (Allen & Yen, 1979). A popular method used to determine construct validity is through

multitrait-multimethod validity (Campbell & Fiske, 1959). Two major types of validity may be determined using this method--convergent validity and discriminant validity. Convergent validity is present when high correlations exist between scores measuring the same trait by different methods; e.g., formal operations task scores measured by paper-and-pencil tests versus Piagetian interview. Notice this is the same way criterion-related validity is determined for paper-and-pencil formal operations tests. Discriminant validity is present when low correlations exist between scores on tests measuring different traits with the same methods; e.g., paper-and-pencil test scores for formal operations tasks and paper-and-pencil test scores of verbal intelligence.

Arlin (1981) developed a formal operations test that was tested for construct validity. The conclusion that there was good construct validity, however, is disputable. Most correlations for convergent validity ranged from .60 to .65 with one correlation of .55 and another correlation of .74. These correlations are minimally acceptable. The discriminant validity coefficients are, with only a few exceptions, below the correlations of convergent validity, but a number of the correlations fall close to them. Staver and Gabel (1979) also found their formal operational test to be lacking in construct validity. Some of the discriminant validity coefficients were actually higher than the convergent

validity coefficients. Bart (1972) found similar results with his formal operations instruments. He found convergent validity coefficients which were significant and positive, but once again, the correlations were modest at best. Further, he found his instruments to have little if any discriminant validity. The formal instruments developed by Bart did not have low correlations with a measure of a different trait. Indeed, the correlations were as high or higher than the correlations determining convergent validity.

This section has painted a bleak picture of paper-and-pencil formal operations tests. The purpose of this section is to point out some measurement issues which need to be considered before formal operations test development progresses further. For example, if formal operations tests are indeed more related to criterion-referenced tests, a reliability index of criterion-referenced tests appropriate to formal operations tests needs to be selected. Berk (1980, 1982) discusses several reliability indices appropriate to criterion-referenced tests which researchers would find useful to consult.

Other Considerations

Two specific problems involved in the issue of interview vs. written formats are response constraint and reading level. Caution needs to be taken when writing items to prevent reading ability from becoming a factor in test response. Blake (1980)

found high school students in Australia who were better readers had an advantage over poor readers when development was assessed by a paper-and-pencil developmental test rather than by the Piagetian interview method. It is intuitively appealing to think reading ability is in someway related to cognitive development, but this is another research question in itself. If reading level is not related to cognitive development, the chance of misclassifying students as formal operational increases.

Response constraint is another issue in paper-and-pencil formal operations tests. Piaget considered the subject's spontaneous response as a crucial factor is his research methodology. Response constraint is not a problem in the clinical interview method because individuals are free to respond at any point in the procedure. Paper-and-pencil tests, with short answer and multiple choice formats, limit subject response. This factor alone leads many to question the appropriateness of paper-and-pencil tests (Elkind, 1971; Herron, 1978). One popular compromise is to allow students to generate their own reasons for the answers they give. While it allows greater freedom of response, it does not allow for interaction between subject, materials, and experimenter, and still leaves scoring up to the subjective interpretation of the scorer.

Summary

This paper has attempted to present some of the problems and issues in the measurement of formal operations. Specifically, an attempt was made to look at the problems with Piagetian interview methodology, solutions to those problems via a paper-and-pencil instrument, and problems with paper-and-pencil formal operations tests. Problems with Piagetian interview methodology include its time consuming procedures and its unstandardized, and to some extent, subjective nature. Paper-and-pencil formal operations tests attempt to solve these problems. They are easily administered to large groups in an efficient and standardized manner. However, formal operations tests present unique problems of their own. Different variables may be tapped by the two methodologies. Further, the methods used to analyze the tests are often inappropriate.

Several conclusions are apparent from the present discussion.

1) Content domains and problem type (e.g., proportional vs. combinatorial) need to be specified and adequately tested before inclusion in formal operations tests.
2) A uniform scoring methodology needs to be developed to increase the interpretability of tests.
3) Validity measures of formal operations tests need to be determined by methods other than just the Piagetian interview. The unstandardized nature of the interview

method threatens the stability of the validity coefficient from one situation to the next.

4) Analyses of formal operations tests should utilize criterion-referenced indices and not those normally used with norm-referenced tests. Formal operations tests are not interested in how formal one subject is compared to all others, but whether a subject is or is not formal; i.e., the interest lies in the precision of our decision and this requires different mathematical formulations (Mehrens & Lehmann, 1975).

It is obvious that this paper offers few, if any, solutions to the problems offered by formal operations measurement. Clarification of what is being measured, i.e, a better definition and description of formal thought, needs to be offered by developmentalists before measurement specialists can solve the reliability and validity problems inherent in the paper-and-pencil test. When evaluating the results of these tests the issues of compatibility with Piagetian techniques, response quality, and domain specificity need to be carefully reviewed as well.

Bibliography

Allen, M. J., & Yen, W. M. (1979). Introduction to Measurement Theory. Monterey: Brook/Cole Publishing Company.

Arlin, P. K. (1982). A multitrait-multimethod validity study of a test of formal reasoning. Educational and Psychological Measurement, 42, 1077-1088.

Bart, W. M. (1972). Construction and validation of formal reasoning instruments. Psychological Reports, 30, 663-670.

Berk, R. A. (1980). A consumers' guide to criterion-referenced test reliability. Journal of Educational Measurement, 17 (4), 323-349.

Berk, R. (1982). Criterion-referenced Measurement. Baltimore: The Johns Hopkins University Press.

Blake, A. J. D. (1980). The predictive power of two written tests of Piagetian developmental level. Journal of Research in Science Teaching, 17, 435-441.

Brainerd, C. J. (1978). Piaget's Theory of Intelligence. England Cliffs, N. J.: Prentice-Hall, Inc.

Breger, L. (1974). From Instinct to Identity. Englewood Cliffs: Prentice-Hall, Inc.

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

Elkind, D. (1981). Children and Adolescents (3rd ed.). New York: Oxford University Press.

Gray, W. M. (1978). A comparison of Piagetian theory and criterion-referenced measurement. Review of Educational Research, 48, 223-249.

Higgins-Trenk, A. & Gaite, A. J. H. (1971). Elusiveness of formal operational thought in adolescents. Proceedings of the 79th Annual Convention of the American Psychological Association, 201-202.

Hunt, J. M. (1975). Psychological development and the educational enterprise. Educational Theory, 25, 333-353.

Lawson, A. E. (1978). The development and validation of a classroom test of formal reasoning. Journal of Research in Science Teaching, 15, 11-24.

Lawson, A. E. (1980). Relationships among level of intellectual development, cognitive style, and grades in a college biology course. Science Education, 64, 95-102.

Lunzer, E. A. (1965). Problems of formal reasoning in test situations. In P. H. Mussen (ed.) Monographs of the Society for Research in Child Development: European Research in Cognitive Development, 30 (2), 19-46.

Martorano, S. (1977). A developmental analysis of performance on Piaget's formal operations tasks. Developmental Psychology, 13, 166-672.

Mehrens, W. A. & Lehmann, I. J. (1978). Measurement and Evaluation in Education and Psychology (2nd ed.). New York: Holt, Rinehart, and Winston.

Modgil, S. & Modgil, C. (1976). Piagetian Research: Compilation and Commentary, v. 4, NFER Publishing Company, Ltd.

Nagy, P. & Griffiths, A. K. (1982). Limitations of recent research relating Piaget's theory to adolescent thought. Review of Educational Research, 52 (4), 513-556.

Piaget, J. (1972). Intellectual evolution from adolescence to adulthood. Human Development, 15, 1-12.

Rowell, J. A. & Hoffman, P. J. (1975). Group tests for distinguishing formal from concrete thinkers. Journal of Research in Science Teaching, 12, 157-164.

Staver, J. R. & Gabel, D. L. (1979). The development and construct validation of a group administered test of formal thought. Journal of Research in Science Teaching, 16, 535-544.

Tobin, K. G. & Capie, W. (1981). The development and validation of a group test of logical thinking. Educational and Psychological Measurement, 41, 413-423.

Tschopp, J. K. & Kurdek, L. A. (1981). An assessment of the relation between traditional and paper-and-pencil formal operations tasks. Journal of Research in Science Teaching, 18 (1), 87-91.

Uzgiris, I. C. & Hunt, J. McV. (1975). Assessment in Infancy: Ordinal Scales of Psychological Measurement. Urbana: University of Illinois Press.

Wollman, W. & Karplus, R. (1974). Intellectual development beyond elementary school, V: Using ratio in different tasks. School Science and Mathematics, 74, 593-613.

Author Notes

The authors wish to thank Toni Santmire and Dave Moshman for their comments and a special thanks to Barbara Plake for her guidance, support, and suggestions on earlier drafts of this article.

Requests for reprints should be sent to Randy Alliger, Department of Educational Psychology and Social Foundations, Teachers College 26, University of Nebraska-Lincoln, Lincoln, Nebraska 68588.